DOCUMENT RESUME

ED 427 070 TM 029 451

AUTHOR Roberson, Thelma J.
TITLE Classroom Observation: Issues Regarding Validity and Reliability.
PUB DATE 1998-11-06
NOTE 25p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (27th, New Orleans, LA, November 4-6, 1998).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; *Data Collection; *Educational Research; Evaluation Methods; *Reliability; Scores; *Teacher Evaluation; *Validity

ABSTRACT

Classroom observation is one of the premiere data collection methods available to those interested in teaching behavior. Observational techniques can be classified on a continuum ranging from low inference to high inference depending on the level of judgment required by the observer making the observation. Central to the issue of any form of measurement are score reliability and validity. This paper explores various observational methods and discusses related reliability and validity issues. These include observer biases, intrusiveness of the observer, observer training, and contextual issues. Validity concerns explored include those of face, content, construct, predictive, and observer validity. Some of the key studies that use classroom observation instruments are described. Further research is suggested based on the "training" factor relevant to the classroom observation. (Contains 31 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 427 070

Classroom Observation: Issues Regarding Validity and Reliability

by
Thelma J. Roberson

The University of Southern Mississippi
Hattiesburg, Mississippi

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the
Mid-South Education Research Association
November 6, 1998
New Orleans, LA

TM029451

BEST COPY AVAILABLE

Abstract

Classroom observation is one of the premiere data collection methods available to those interested in teaching behavior. Observational techniques can be classified on a continuum ranging from low inference to high inference depending on the level of judgment required by the observer making the observation. Central to the issue of any form of measurement is that of score reliability and validity. This paper explores various observational methods and discusses related validity and reliability issues. The paper suggests further research based on the "training" factor relevant to the classroom observation.

Classroom Observation

Traditionally, *classroom observation* has been the foremost method for gathering data regarding teaching and teacher behavior. It involves the systematic matriculation of specified behavior that is postulated, in light of a theoretical framework, to be commensurate with positive student development (McGreal, 1983). Classroom observation is an intentional, methodical process that is planned and focused. It involves more than merely "seeing", it requires the full attention of the observer (Hyman, 1975) and the ability to properly record what has occurred in the observational setting.

The importance and use of classroom observation has placed it in a genera viewed by some as the most pragmatic procedure for collecting formal data about teacher performance (McGreal, 1983) and thus its usefulness as an evaluative tool has been propagated. Because observation has the capacity to disclose the climate, compatibility, interactions, and operations of the classroom which is available from no other source (Evertson & Holley, 1981) observational assessment is frequently used singularly or with other assessment techniques to comprise a teacher's summative evaluation (Elliot & Mattar, 1990).

Madeline Hunter (1988) considers classroom observation as the "...most heavily weighted evidence of professional excellence..." (p.53) and asserts that proof of possession of certain teaching competencies can be validated through the use of recurring observations. Joyce and Showers (Costa, et al., 1988, p. 145) state, "...we have become convinced that the overt, [observable] skills (of teaching) are driven by mental activities that constitute the invisible skills of teaching." Hyman (1975) proclaims that observation is the only means of gathering evidence needed

for the evaluation process. These and similar prevailing thoughts have perpetuated the use of classroom observation and rating scales as teacher evaluation tools since the early 1900s. So much so, to question it "...smacks of heresy" (Medley, 1982). However, if observation is to occupy such a prominent position in the educational arena, its validity and reliability must be ascertained. The purpose of this paper is to explore the concepts of classroom observation and question its validity and reliability.

Types of Observations

Observational techniques variegate based on a variety of factors and may range from a wide lens, anecdotal record style approach which attempts to record every incident (Acheson & Gall, 1997) to a very narrow, selective rating system that excludes behaviors not specifically identified by the process (Medley, 1982). When used as an evaluative tool, as is frequently the case, observations require varying degrees of value judgments. For a value judgment to be considered acceptable, it must be supported by evidence and relevant criteria (McGreal, 1983).

Low Inference Observations

An observation is considered *low inference* when the behaviors to be observed are specifically prescribed and predetermined before the observation takes place. The role of the observer is to be a collector of descriptive data (McGreal, 1983) who simply looks for these well-defined behaviors and decides if they are or are not occurring during the observation period. No part of the observation requires the observer to make qualitative judgments (Wiersma, et al. 1983) as all inferences involved occurred beforehand in the selection of criteria and in the development of the scoring key (Medley et al. 1984). McGreal (1983) contends

that it is only through this type of selective surveillance that an observation can approximate any degree of reliability due to the complex nature of the classroom arena.

While low inference observations appear to be highly objective, the problem lies in the fact that one's philosophical orientation may 'cloud' how a behavior is viewed (Medley, et al. 1984). "Being selective involves taking a point of view..." (Hyman, 1975, p. 25); therefore, the benefits of selectivity force observers to become somewhat subjective. Where one observer views a student's behavior during an observation as a display of independent thinking, another may view the same behavior as a classroom disruption. A third observer may find the behavior irrelevant and not identify the behavior at all. This difference in philosophic inclination could therefore result in three different scores for the same behavior (Medley, et al. 1984).

High Inference Observations

When an observation requires a substantial degree of inferences or qualitative judgments on the part of the observer, it is considered a *high inference observation*. Judgments may be necessary to determine if a particular behavior is actually occurring and if so, to what degree and with what merit. Often, the observer must decide what numerical value to assign observed behaviors (Wiersma, et al., 1983). Medley, et al. (1984) views such high inference observations with skepticism:

> If decisions about relevance and weighting are difficult for you to make when you are developing a scoring key and have ample time to deliberate and consult the literature and your colleagues, how can you expect your rater to make them wisely under the conditions in

> which he works? ...the task of recording behavior accurately and objectively is about all an observer can handle successful. (p. 44)

Although the process is highly subjective by nature, it has the advantage of providing a scoring mechanism that closely approximates an interval scale which enables comparative analysis helpful in ranking teachers and establishing reliability (Wiersma, et al. 1983). Another advantage of the high inference observation is its flexibility. Because environmental context is a critical part of behavior formation, having too narrow a focus may limit the observer's ability to capture the essence of the classroom experience, thereby limiting the quality of the observation's results. Teacher behavior does not occur in a vacuum but is partly a response to the interplays of the classroom setting (Evertson & Holley, 1981). High inference observation allows the observer more of an opportunity to interpret these interplays.

Observation Instrument

Since 1915, rating scales and observational instruments have abounded. By 1930, over 200 had been identified (Medley, 1982) and were in use. Many of these exist today either in original or revised forms and numerous instruments have been added creating a wealth of resources from which to choose. The advantage to selecting a constructed observational instrument is that it contains a built-in framework, a vantage point, a philosophical orientation. Existing instruments also include a set of rules for systemic observations and data compilation and generally have some measure of reliability (Hyman, 1975). Poster and Poster (1991) caution though that while instruments with established performance criteria are helpful in establishing norms of behavior, they must be moderated by the prevailing circumstances of the individual educational setting. The process is basically reduced to the selection of an instrument that reflects the needs and beliefs of the organization and is

within its financial and personnel capabilities. Should an organization choose to develop its own observation instrument, there are numerous resources available to guide in its creation (Acheson & Gall, 1997, Peterson et al., 1985, & Stanley & Popham, 1988).

Regardless of the instrument chosen, it must be realized that an observational situation is a testing situation. Valid results can only be obtained when the teacher performs at optimum levels (Medley, et al., 1984). Therefore, it is imperative that the observer notify the teacher of the instrument that will be used during the observation period (Hyman, 1975). While there are conflicting views on whether or not an observation should be announced or unannounced, Madeline Hunter (1988) points out that:

> ...an effective teacher will not teach poorly because they did not expect an observation. An ineffective teacher will not magically develop (teaching skills) the night before... (p. 46)

In light of this view, choosing to announce or not announce an observation seems somewhat irrelevant if the observation is a reflection of the teachers capabilities.

It should be apparent that classroom observation instruments contain a large gamut of characteristics and differ in scope, focus, direction, uses, and results. Their commonality seems to be in the necessity and difficulty in establishing and maintaining their objectivity and reliability (Elliot & Mattar, 1990).

Observational Inconsistencies

Errors and inconsistencies occurring with the use of classroom observation techniques can be classified into several categories including observer biases (Boehm & Weinberg, 1987, Manatt, 1988, Gronlund, 1971, McGreal, 1983; Medley 1982,

& Medley et al., 1984), intrusiveness (Jeager, 1993 & Medley, et al., 1984), training (Boehem & Weinberg, 1987 & Manatt, 1988), and contextual issues (Hunter, 1988, & Popham, 1988).

Observer Biases

Scoring tendency

One form of observer bias results from the *scoring tendency* (Manatt, 1988) of the observer. Some observers tend to score teachers in a noticeable pattern which may be lenient, severe, or toward the middle (central tendency). Scoring patterns have been identified, in a very liberal sense, with the following groups of observers:

- Females tend to rate male and female teachers more severely than do their male counterparts.
- Hispanic observers rate teachers the lowest of all ethnic groups. Black observers tend to give the next to the lowest ratings.
- Observers with higher levels of education (above a master's degree) tend to be more lenient in rating teachers. Scores tend to go up proportionately to the observer's level of education.
- The more experienced an observer becomes, the more likely a tendency develops to score teachers more leniently.
- The position of an observer tends to alter rating patterns. The greater the status, the more severe the ratings. Example, superintendents often feel principals are too lenient with teacher ratings.

Halo Effect / Prejudice

Another form of observation error results from the *halo effect* (Gronlund, 1971, Manatt, 1988, Boehm & Weinberg, 1987, & Medley et al., 1984). When an observer holds an overall positive impression of a teacher, there is a tendency to

rate that teacher favorably whether or not there is evidence that the teacher is actually displaying the desired behavior (Medley et al., 1984). When just the opposite occurs and a teacher is scored negatively based on the observers overall impression, the unfortunate state is considered *prejudice* and is another source of observation error (Boehm & Weinberg, 1987). Some prejudices include not only traditionally recognized injustices such as race and gender biases but include theoretical and philosophical differences as well.

Logical Error

Gronlund (1971) pointed out that observer error can occur when two characteristics are assumed to have some relationship to each other and therefore there exists the belief that a teacher's score should reflect similar ratings on each. One example of this would be the assumption that the characteristics of intelligence and achievement have a significant relationship with each other and therefore, the scores in one area should be reflective of scores in the other. This misconception known as *logical error* effects scores similarly to the halo effect. A score on a given characteristic is either inflated or deflated based on its perceived relationship to another characteristic and its score.

Observer Drift

Observer drift is another root of observation error. This phenomenon begins to occur when an observer has become seasoned or highly experienced or when a pair of observers have worked together for a period of time. The latter is referred to as *consensual drift*. Boehm and Weinberg (1987) characterize a drift as evidenced by a lessening in the precision and accuracy of observations. The observer becomes somewhat desensitized to the classroom environment and does not focus as well as

with previous observations. When pairs of observers work together for a period of time, scores tend to become more alike and there is less discussion justifying the appropriateness of given scores.

Intrusiveness / Heinsenberg Principle

The presence of a classroom observer is often intrusive, altering the natural flow of classroom activities, and leads to a phenomenon referred to as the *Heisenberg Principle*. This occurs when the act of measurement alters what is being measured. While use of a video camera may lessen the effect, it too is intrusive (Jaeger, 1993) and is limited by its technological capabilities. Intrusions of any sort lead to a distortion of the classroom environment (Popham, 1988) so the activities and interactions are no longer typical. When observations report atypical behaviors, results are limited and possibly invalid. The degree to which intrusions actually affect observation error is abstrusely difficult to calculate because as Medley, et al. (1984, p. 135) points out "...very little empirical research is produced ...because it is so difficult to get reliable observations of what happens in the classroom when no one is there to observe."

Inadequate Observer Training

One of the leading causes of error, and one that will be discussed further in another section, is the insufficiency of competent observer training. Boehm & Weinberg (1987) contend that an observer is not adequately trained until reaching a 90% agreement rate with a master observer during training sessions. While the amount and types of training recommended by researchers and instrument developers varies remarkably, the consistent factor is that all observers need to be trained. Manatt (1988) proposes that training is the key ingredient to reducing the observational errors mentioned heretofore.

Complexity of Educational Setting / Teacher Proficiency Level

Other sources of error reflect the complexity of the classroom environment and the proficiency levels of the teachers being observed. It is ironious to think that one instrument could address all the possible contextual and experiential situations that could be present in any given educational setting. Berliner (1990) identified at least five stages of teacher proficiency ranging from novice to expert and was circumspect in perceiving the highly contextual nature of the identification of each. An 'expert' teacher in one setting may not be so in another. Observers must recognize the context and set of circumstances surrounding a teaching situation. For example, an observer's perception of classroom activity can be greatly altered by the simple knowledge of where the group is in regards to the lesson. Is this the beginning, middle, or end of a unit of study? Due to the contextual issues surrounding every teaching situation, no formal observation should take place without a prior conference (McGreal, 1983) and an opportunity for the observer to become somewhat oriented to the teaching environment. In an evaluative sense, the observation criteria must be appropriate for the skill and maturity of the teacher (Hunter, 1988) and as the National Education Association proposes, "...under no circumstances should a teacher be evaluated except by considering elements of the teacher's specific instructional situation" (Popham, 1988, p. 69).

Legal Considerations

When observational methods are used solely or in conjunction with other assessment measures as a summative evaluation tool, they must meet the same legal guidelines as competency and employability tests. Legal precedent holds all employment testing to the standards set forth by the Equal Employment Opportunity Commission and the American Psychological Association in regards to nondiscriminatory employment practices. This includes all competency testing and observational

assessments relative to employment decisions. Recent court rulings in these areas have tended to uphold the usability of such measures when has been shown that a maximum effort was made to avoid biases and where there was minimal impact to minority employees (Rebell, 1990). These legal implications are mentioned here because more and more states, especially in the southeastern region (Ellet, et al., 1994), are adopting mandatory statewide evaluation systems which include some form of classroom observation. Those who develop and use such instruments must recognize the legalities involved and make efforts to ensure that the individual rights of those being evaluated are protected.

Validity

is an instrument's ability to measure what it is intended to measure (Daniel & Siders, 1994), the degree to which inferences from obtained scores are supported by evidence (Elliot & Mattar, 1990), or as Manatt (1988)simply put it, validity is truthfulness. There are at least five different types of validity. While all have merit, some are more meaningful to classroom observation than others.

Face Validity

The degree to which an instrument 'appears' to be adequate refers to its *face validity*. Because an observational instrument may include items that 'appear' to model good teaching strategies that may or may not have any research base to support them, face validity is not typically a sound psychometric measure. It does, however, serve a valid screening purpose (Gay, 1996). Daniel and Siders (1994) warn that face validity must be interpreted with extreme caution and that its true value lies only in the formation of hypothesizes regarding the correlation of criteria and observational instruments.

Content Validity

Content validity refers to the sampling of the items contained in an instrument (Evertson & Holley, 1981). It asks, do these items exemplify a logical sampling of items indicative of the trait being measured? Nunnally (Daniel & Siders, 1994) emphasizes that content validity relies heavily on reason to determine the adequacy of the sampling. This is true because there is no computational method for establishing content validity and the adequacy of its sampling can only be determined by expert judgment (Gay, 1996).

Construct Validity

Construct validity deals with traits not readily observable (Evertson & Holley, 1981). Constructs are the abstract conceptions that underlie the variables. Nunnally (Daniel & Siders, 1994) equates construct validity to factorial or trait validity. For each teacher trait to be accurately identified, there must be an underlying hypothesis as to how a person with such a trait would behave in a given situation. Therefore, establishing construct validity is actually the testing of these underlying hypotheses (Gay, 1994).

Predictive Validity

When the results from one set of observational data are predictive of another set of data the instrument used during the observation is said to have *predictive validity* (Evertson & Holley, 1981). As the purpose of classroom observation is to evaluate and improve the educational setting it is clear that the "...main ingredients required for validation studies are valid measures of both teacher performance and student outcomes" (Peterson et al., 1985, p. 170), hence, observational measures that contain predictive validity. However, predictive validity is often extremely difficult to measure because it requires comparable data and students' standardized test scores

are generally nontransferable

Observer Validity

While some regard observer agreement as a reliability measure, Medley, et al. (1984) contend that this is a misconception. They contend that observer agreement is actually *observer validity* because it measures the degree to which the frequencies of the scores from an observation agree with what is actually occurring. Observer validity is heavily dependent upon several key factors including: observer training, clarity of items and their definitions, internal consistency of the instrument's scoring key, differences among individual teachers, and the stability of the behaviors being measured.

Regarding the overall issue of validity, Justice O'Conner wrote, "... tests (including observational data) are not valid or invalid per se, but must be evaluated in the setting in which they are used; the fact that the validity of a particular (instrument) has been ruled upon in prior litigation is not necessarily determinative in a different factual setting" (Rossow & Parkinson, 1992). This concept is expanded upon by Croll (1986) who states "... all social research involves a purposive abstraction from social phenomena and that the validity of descriptions are limited by their purposes" (p. 155).

In general validity should be established through numerous observations and should continue to be checked as new objectives are added (Medley, et al., 1984). Although there is an emerging body of knowledge regarding effective teaching, (Evertson & Holley, 1981) validity will remain suspect until an agreed upon

set of standards exists. As there is presently no clear definition of what constitutes effective teaching, there is also no current defensible means of teacher evaluation (Mehrens, 1987). Dwyer (Shinkfield & Stufflebeam, 1995) feels it is this absence of educationally, technologically, logically, and ethically defensible criteria of what effective teaching is that leaves observation and other assessment areas open to criticism and its validity in doubt. Medley, et al. (1984) cite an example of the misconceptions regarding effective teaching criteria. Most assessment instruments include measures of how a teacher individualizes instruction yet in at least one study it was found that small group and individualized instruction was actually associated with less learning.

Reliability

Evertson and Holley (1981) assert that observations can be reliable. That is, they can produce the same results on subsequent occasions (Croll, 1986). Reliability identifies the degree of measurement error (Evertson & Holley, 1981) of a given indicator or instrument by identifying how consistently (Elliot & Mattar, 1990) like results are obtained. In essence, reliability is consistency (Manatt, 1988).

One of the main problems encountered when attempting to establish the reliability of observational data is the confusion regarding what the term reliability means in this context (Croll, 1986). Just as face validity must not be confused with predictive validity, some research methods are more appropriate than others in determining reliability. As noted earlier, Medley, et al. (1984) categorize observer agreement as a measure of validity, yet some attempt to use this measure to establish the reliability of an instrument. When this occurs, reliability estimates are inflated and invalid results are obtained. A more appropriate study would investigate the reliability coefficient, that is the degree to which individuals can consistently be ranked (Croll, 1986)

when observed by different observers on different occasions and even with different instruments (Medley, et al., 1984). Croll, (1986) refers to the reliability coefficient as the stability measure.

Reliability should be considered as a matter of degree (Medley, et al., 1984) rather than an absolute. Reliability studies must consider that reliability measures are highly dependent on three main factors: the internal consistency of the instrument's scoring key, the accuracy of the observer, and the stability of the behaviors being measured (Medley, et al., 1984). The more sufficient the factors are, the greater the degree of reliability.

Key Studies

While there appears to be a dearth in the area of empirical research establishing the validity and reliability of classroom observation instruments, there have been a few key studies that have broached the subject. The work of Medley, et al. (1984) found through an exhaustive review of the literature that observation methods do not serve predictive functions in regards to teacher effectiveness and therefore can not be considered valid measures. Their research of various rating scales revealed that teachers with high ratings were no more effective than teachers with low ratings and there was no relationship between teacher scores and pupil learning.

Weirsma, et al. (1983) made a comparative study of low and high inference instruments using the COKER (Classroom Keyed for Effectiveness Research), a low inference instrument and the TPAI (Teacher Performance Assessment Instrument), a high inference instrument. Through a process of mapping, 18 components were identified that were measured by both instruments and were used in the study. The results showed little convergent validity and the results were interpreted to indicate

that measurement is highly instrument dependent. The study also found that a single factor accounted for eight times as much variance as did any of the other 17 factors. Medley, et al. (1984) asserted that this factor was actually measuring the halo effect.

Most studies have targeted specific instruments that include classroom observation as an evaluative tool rather than looking at classroom observation as a whole. While getting mixed results, some of these studies have established varying degrees of validity and reliability for individual instruments.

MTAI

The *MTAI* (Mississippi Teacher Assessment Instrument), once mandated as a certification requirement for all Mississippi teachers, is heavily reliant upon the classroom observations of trained observers. Daniel and Siders (1994) utilized a factorial analysis approach and found the instrument to contain some degree of construct validity. A previous study conducted by Siders, et al. had established the instrument's content validity. What the instrument lacks is predictive validity as evidenced by the work of its developers and through a comparative study using NTE (National Teacher's Exam) subtest scores.

FPMS

Studies of the Florida Performance Measurement System (*FPMS*) revealed evidence of an acceptable level of validity and reliability. This is not surprising considering the intensive research efforts that went into the development of the instrument. The FPMS is one of few instruments that has established norms whereby teachers can be compared with standardized results. No doubt this feature has aided in the establishment of its validity and reliability measures (Peterson, et al., 1985).

STAR

Several validity and reliability studies have been conducted with the Louisiana's *STAR* (System for Teaching and Learning Assessment Review) evaluation program (Ellett, 1991, & Chauvin, 1991). These studies found STAR results to be valid and reliable indicators of effective teaching with inferences to pupil learning. The recent work of Ellet, et al. (1994) confirmed the instrument's construct validity but raised serious questions regarding the validation process researchers use to establish construct validity.

Discussion

With so many varying forms of classroom observation, its highly contextual nature, and with so much inconclusive evidence surrounding its validity and reliability, it would seem that researchers would look for some commonalities upon which to base future research efforts. Rather than debating about which instruments are valid or how one instrument compares to another, why not look at the key ingredient that makes auspicious instruments successful and to what degree that 'ingredient' can affect the validity and reliability of its results. The one prominent resounding component common to all observation instruments is the training of its observers.

As Bitner and Kratzner (1995) maintain, no observation can be reliable without the prior adequate training of the observer. This indisputable fact and the effects that observational inconsistencies have on the results of the observation should provide the basis for inquiry into the effects that training plays on the validity and reliability of any observational instrument.

In state mandated evaluation systems, it was found that over 80% of classroom observations were being conducted by line administrators (McGreal, 1983). The range of experience and training each has received in observation techniques and utilization of the instrument being used varies remarkably. Madeline Hunter (1988) contended that observer training should consist of no less than 50-100 hours under the tutelege of a proficient trainer. To suggest self-study as an apropos training method is as inappropriate as assuming competent surgeons or accomplished musicians could do the same to develop their skills. Yet, this is how some observers have been trained, reading the manual. Some instrument developers recommend prescribed training sessions. The STAR's manual for instance suggests several weeks of extensive training (Medley, et al., 1984). Others require only two hours (McGreal, 1983).

Since observer training has already been identified as a redress to observation inconsistencies (Boehm & Weinberg, 1987, & Manatt, 1988), it would seem wise then to further investigate its effects upon the instrument being studied. Training may be the most determinant variable when measuring the validity and reliability of adequate observation instruments. It is a training's ability to instill professional expertise within the observer that is important. For with this expertise, "...a trained observer should be able to distinguish quality when observing the teaching and learning situation" (Wiersma, et al., 1983) therefore, valid results can be obtained. Without the expertise to use an instrument for its intended purposes, it is an affront to claim that the results are valid. Cangelosi (1991) addresses the matter by stating:

> ...unless extensive efforts are undertaken to educate and train observers to use well-designed observational instruments, classroom observations will continue to be dominated by malpractice that produces invalid results. (p. 47).

It is for this reason that researchers should turn their attention to the training component of classroom observations. Their findings will undoubtedly become the driving force behind impending observation instrument training and future validity studies in this area.

References

Acheson, K. A., & Gall, M. D. (1997). Techniques in the clinical supervision of teachers: Preservice and inservice applications (4th ed.). White Plains, NY: Longman Publishers USA.

Berliner, D. C. (1990). Implications of studies of expertise in pedagogy for teacher education and evaluation. In The assessment of teaching: Selected topics (pp. 21-50). Amherst, MA: National Evaluation Systems, Inc.

Bitner, T. & Kratzner, R. (1995). A primer on building teacher evaluation instruments. Chicago, IL: Midwest Educational Research Association. (ERIC Document Reproduction Service No. ED 394 953)

Boehm, A. E., & Weinberg, R. A. (1987). The classroom observer: Developing observation skills in early childhood settings (2nd ed.). New York, NY: Teacher's College Press.

Cangelosi, J. S. (1991). Evaluating classroom instruction. White Plains, NY: Longman Publishing Group.

Chauvin, S. W. (1991). Development and validation of a comprehensive assessment system for teaching and learning. Chicago, IL: National Counsel on Measurement in Education. (ERIC Document Reproduction Service No. ED 335 410).

Costa, A. L., Garmston, R. J., & Lambert, L. (1988). Evaluation of teaching: The cognitive development view. In R. S. Brandt (Series Ed.) & S. J. Stanley & W. J. Popham (Vol. Eds.). Teacher evaluation: Six prescriptions for success (pp. 145-172). Alexandria, VA: Association for Supervision and Curriculum Development.

Croll, P. (1986). Systematic classroom observation: Social research and educational studies series: 3. Philadelphia, PA: The Falmer Press, Taylor & Francis, Inc.

Daniel, L. G., & Siders, J. A. (1994). Validation of teacher assessment instruments: A confirmatory factor analysis approach. Journal of personnel evaluation in education, 8 (1), 29-40.

Elliot, S. M., & Mattar, J. D. (1990). Beyond traditional assessment methods: Alternative approaches for assessing entry-level teachers. In The assessment of teaching: Selected topics (pp. 51-64). Amherst, MA: National Evaluation Systems, Inc.

Ellett, C. D. (1991). Development, validity, and reliability of a new generation of assessments of effective teaching and learning: Future directions for the study of learning environments. Journal of classroom interaction, 26 (2), 25-36.

Ellett, C. D., Loup, K. S., Evans, R. L., Chauvin, S. W., & Naik, N. S. (1994). A study of teachers' nominations of superior colleagues: Implications for teacher evaluation programs and the construct validity of classroom-based assessments of teaching and learning. Journal of personnel evaluation in education, 8 (1), 7-28.

Evertson, C. M., & Holley, F. M. (1981). Classroom observation. In J. Millman (Ed.), Handbook of teacher evaluation (pp. 90-109). Beverly Hills, CA: SAGE Publications.

Gay, L. R. (1996). Educational research: Competencies for analysis and application (5th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Grolund, N. E. (1971). Measurement and evaluation in teaching (2nd ed.). New York, NY: Macmillan Company.

Hunter, M. (1988). Create rather than await your fate in teacher evaluation. In R. S. Brandt (Series Ed.) & S. J. Stanley & W. J. Popham (Vol. Eds.). Teacher evaluation: Six prescriptions for success (pp. 32-54). Alexandria, VA: Association for Supervision and Curriculum Development.

Hyman, R. T. (1975). School administrator's handbook of teacher supervision and evaluation methods. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Jaeger, R. M. (1993). Live vs. memorex: Psychometric and practical issues in the collection of data on teachers' performances in the classroom. Atlanta, GA: American Educational Research Association. (ERIC Document Reproduction Service No. ED 360 325)

Manatt, R. P. (1988). Teacher performance evaluation: A total systems approach. In R. S. Brandt (Series Ed.) & S. J. Stanley & W. J. Popham (Vol. Eds.). Teacher evaluation: Six prescriptions for success (pp. 79-108). Alexandria, VA: Association for Supervision and Curriculum Development.

McGreal, T. L. (1983). Successful teacher evaluation. Alexandria, VA: Association for Supervision and Curriculum Development.

Medley, D. M. (1982). Teacher competency testing and the teacher educator. Charlottesville, VA: Bureau of Educational Research.

Medley, D. M., Coker, H., & Soar, R. S. (1984). Measurement-based evaluation of teacher performance: An empirical approach. New York, NY: Longman, Inc.

Mehrens, W. A. (1987). Issues in teacher competency tests. Jacksonville, FL: Institute for Student Assessment and Evaluation at the University of Florida. (ERIC Document Research Service No. ED 337 487)

Peterson, D., Micceri, T., & Smith, B. O. (1985). Measurement of teacher performance: A study in instrument development. In L. W. Barber (Series Ed.) & G. C. Hall (Vol. Ed.), Phi Delta Kappa center on evaluation, development and research hot topics series: Teacher competence (pp. 157-171). Bloomington, IN: Phi Delta Kappa.

Popham, W. J. (1988). Judgment-based teacher evaluation. In R. S. Brandt (Series Ed.) & S. J. Stanley & W. J. Popham (Vol. Eds.). Teacher evaluation: Six prescriptions for success (pp. 56-77). Alexandria, VA: Association for Supervision and Curriculum Development.

Poster, C., & Poster, D. (1991). Teacher appraisal: Training and implementation (2nd ed.). New York, NY: Routledge.

Rebell, M. A. (1990). Legal aspects of subjective assessments of teacher competency. In The assessment of teaching: Selected topics (pp. 1-10). Amherst, MA: National Evaluation Systems, Inc.

Rossow, L. F., & Parkinson, J. (1992). The law of teacher evaluation. NOLPE monograph/book series no. 42. Topeka, KS: National Organization on Legal Problems of Education. (ERIC Document Reproduction Service No. ED 337 851)

Shinkfield, A. J., & Stufflebeam, D. (1995). Teacher evaluation: Guide to effective practice. Boston, MA: Kluwer Academic Publishers.

Stanley, S. J., & Popham W. J. (Eds.). (1988). Teacher evaluation: Six prescriptions for success. Alexandria, VA: Association for Supervision and Curriculum Development.

Wiersma, W., Dixon, G. E., Jures, S., & Wenig, J. (1983). Assessment of teacher performance: Constructs of teacher competencies based on factor analysis of observation data. Montreal, Quebec: American Educational Research Association. (ERIC Document Reproduction Service No. ED 230 586)